LUCAS AGAINST MECHANISM

DAVID LEWIS

J. R. LUCAS argues in "Minds, Machines, and Gödel",[1] that his potential output of truths of arithmetic cannot be duplicated by any Turing machine, and *a fortiori* cannot be duplicated by any machine. Given any Turing machine that generates a sequence of truths of arithmetic, Lucas can produce as true some sentence of arithmetic that the machine will never generate. Therefore Lucas is no machine.

I believe Lucas's critics have missed something true and important in his argument. I shall restate the argument in order to show this. Then I shall try to show how we may avoid the anti-mechanistic conclusion of restated argument.

As I read Lucas, he is rightly defending the soundness of a certain infinitary rule of inference. Let L be some adequate formalization of the language of arithmetic; henceforth when I speak of sentences, I mean sentences of L, and when I call them true, I mean that they are true on the standard interpretation of L. We can define a certain effective function *Con* from machine tables to sentences, such that we can prove the following by metalinguistic reasoning about L.

C1. Whenever M specifies a machine whose potential output is a set S of sentences, *Con* (M) is true if and only if S is consistent.

C2. Whenever M specifies a machine whose potential output is a set S of true sentences, *Con* (M) is true.

C3. Whenever M specifies a machine whose potential output is a set S of sentences including the Peano axioms, *Con* (M) is provable from S only if S is inconsistent.

Indeed, there are many such functions; let *Con* be any chosen one of them. Call ϕ a *consistency sentence* for S if and only if there is some machine table M such that ϕ is *Con* (M) and S is the potential output of the machine whose table is M. Now I can state the rule R which I take Lucas to be defending.

R. If S is a set of sentences and ϕ is a consistency sentence for S, infer ϕ from S.

Lucas's rule R is a perfectly sound rule of inference: if the premises S are all true, then by C2 so is the conclusion ϕ. To use R is to perform an inference in L, not to ascend to metalinguistic reasoning about L. (It takes metalinguistic reasoning to show that R is truth-preserving, but it takes metalinguistic reasoning to show that *any* rule is truth-preserving.)

Lucas, like the rest of us, begins by accepting the Peano axioms for arithmetic. (Elementary or higher-order; it will make no difference.) A sentence ψ is a theorem of Peano arithmetic if and only if ψ belongs to every superset of the axioms which is closed under the ordinary rules of logical inference. Likewise, let us say that a sentence χ is a theorem of *Lucas arithmetic* if and only if χ belongs to every superset of the axioms

which is closed under the ordinary rules of logical inference and also closed under Lucas's rule R. We have every bit as much reason to believe that the theorems of Lucas arithmetic are true as we have to believe that the theorems of Peano arithmetic are true: we believe the Peano axioms, and the theorems come from them by demonstrably truth-pre serving rules of inference. Knowing this, Lucas stands ready to produce as true any theorem of Lucas arithmetic.[2]

Suppose Lucas arithmetic were the potential output of some Turing machine. Then it would have a consistency sentence ϕ. Since Lucas arithmetic is closed under R, ϕ would be a theorem of Lucas arithmetic. Then ϕ would, trivially, be provable from Lucas arithmetic. Then, by C3, Lucas arithmetic would be inconsistent. Lucas arithmetic would contain falsehoods, and so would the Peano axioms themselves. Therefore, insofar as we trust the Peano axioms, we know that Lucas arithmetic is not the potential output of any Turing machine. Assuming that any machine can be simulated by a Turing machine—an assumption that can best be taken as a partial explication of Lucas's concept of a machine—we know that neither is it the potential output of any machine. Thus if Lucas arithmetic is the potential output of Lucas, then Lucas is no machine.

So far, so good; but there is one more step. Although Lucas has good reason to believe that all the theorems of Lucas arithmetic are true, it does not yet follow that his potential output is the whole of Lucas arithmetic. He can produce as true any sentence which he can somehow *verify* to be a theorem of Lucas arithmetic. If there are theorems of Lucas arithmetic that Lucas cannot verify to be such, then his potential output falls short of Lucas arithmetic. For all we know, it might be the potential output of a suitable machine. To complete his argument that he is no machine—at least, as I have restated the argument—Lucas must convince us that he has the necessary general ability to verify theoremhood in Lucas arithmetic. If he has that remarkable ability, then he can beat the steam drill—and no wonder. But we are given no reason to think that he does have it.

It is no use appealing to the fact that we can always verify theoremhood in any ordinary axiomatic theory—say, Peano arithmetic—by exhibiting a proof. True, if we waive practical limitations on endurance; but Lucas arithmetic is not like an ordinary axiomatic theory. Its theorems do have proofs; but some of these proofs are transfinite sequences of sentences since Lucas's rule R can take an infinite set S of premises. These transfinite proofs will not be discovered by any finite search, and they cannot be exhibited and checked in any ordinary way. Even the finite proofs in Lucas arithmetic cannot be checked by any mechanical procedure, as proofs in an ordinary axiomatic theory can be. In order to check whether Lucas's rule R has been used correctly, a checking procedure would have to decide whether a given finite set S of sentences was the output of a

machine with a given table M. But a general method for deciding that could easily be converted into a general method for deciding whether any given Turing machine will halt on any given input—and that, we know, is impossible.

We do not know how Lucas verifies theoremhood in Lucas arithmetic, so we do not know how many of its theorems he can produce as true. He can certainly go beyond Peano arithmetic, and he is perfectly justified in claiming the right to do so. But he can go beyond Peano arithmetic and still be a machine, provided that some sort of limitations on his ability to verify theoremhood eventually leave him unable to recognize some theorem of Lucas arithmetic, and hence unwarranted in producing it as true.[3]

University of California at Los Angeles.

[1] *Philosophy*, 36 (1961): 112-127.

[2] Lucas arithmetic belongs to a class of extensions of Peano arithmetic studied by A. M. Turing in "Systems of Logic Based on Ordinals", *Proceedings of the London Mathematical Society*, sec. 2, 45 (1939): 161-228, and by S. Feferman in "Transfinite Recursive Progressions of Axiomatic Theories", *Journal of Symbolic Logic*, 27 (1962): 259-316.

[3] I am indebted to George Boolos and Wilfrid Hodges for valuable criticisms of an earlier version of this paper.